



ABBYY®

FineReader® OCR XIX

Based on FineReader 7.0

Erste Omnifont OCR zur Erkennung von Frakturschrift und alten europäischen Sprachen

ABBYY FineReader XIX ist eine spezielle Version der vielfach ausgezeichneten OCR Software FineReader für die Erkennung von Texten, die zwischen 1800 und 1938 in Frakturschrift gedruckt wurden. ABBYY FineReader XIX ist die erste Omnifont OCR Software am Markt für Frakturschrift.

Die Herausforderung: Digitalisierung alter Texte

Bis heute hatten noch nicht ausgereifte Technologien und die Einzigartigkeit von Texten, die in unterschiedlichen alten Schriftarten verfasst waren, die automatische Erfassung dieser Texte durch Computer erschwert wenn nicht gar unmöglich erscheinen lassen. Anspruchsvolle Wörterbücher und Sprachmodelle, die für die Analyse und Verifizierung der Texte verwendet werden können, gab es bis jetzt noch nicht. Computersysteme, die diese Texte lesen konnten, mussten erst viele Stunden trainiert werden, um Schriftarten und Zeichen zu erkennen, die heute gar nicht mehr verwendet werden.

Gebrochene Schriften entstanden erstmals im 12. Jahrhundert und entwickelten sich im Laufe der Jahrhunderte in vielen verschiedenen Variationen weiter. Die Frakturschrift, vorherrschend in Deutschland, wurde durch Kaiser Maximilian eingeführt und hat sich schnell in vielen Teilen Europas etabliert. Die Besonderheiten dieser Schrift beinhalten das verlängerte "s" und Ligaturen oder zusammenhängende Buchstaben bei bestimmten Kombinationen. Die Erscheinungshäufigkeit dieser Eigenart ist entscheidend für das Verständnis von Frakturschrift, wenn man Erkennungstechnologien für Texte, die aus der Zeit zwischen 1800 und 1938 entstanden, entwickeln möchte.

Die ABBYY Lösung: Erste Omnifont OCR für Frakturschrift

ABBYY FineReader XIX ist die erste Omnifont OCR für Fraktur und somit eine Lösung für Anwender, die alte Texte mit wenig Aufwand scannen und umwandeln möchten. Dies wurde durch Kombination einer besonders intelligenten Erkennungs-Technologie und ausführlichen linguistischen Studien erreicht:

OCR Systeme analysieren Texte und stellen Hypothesen darüber auf, welcher Buchstabe oder welches Wort durch ein Bild dargestellt werden. Diese Hypothesen werden daraufhin im Kontext analysiert und durch den Einsatz anspruchsvoller OCR Wörterbücher, die aus Sprachmodellen bestehen, verifiziert. Diese Sprachmodelle sind komplexe Datenbanken, die das Vokabular einer Sprache beschreiben. Moderne OCR Systeme verfügen jedoch über keine Sprachmodelle für ältere Schriftarten oder Schreibweisen. Nach der ABBYY-eigenen Entwicklung von Modellen für fünf europäische Sprachen speziell für diese Zeit können auch derartige Texte nun verarbeitet werden.

Dabei wurden 10 verschiedene Wörterbücher und mehr als 105 Bücher, die zwischen 1808 und 1930 veröffentlicht wurden, analysiert. Linguisten überprüften Wortstämme, identifizierten Wörter, die bei der Entwicklung der Sprache nicht weitergeführt wurden, und bestimm-

ten die korrekten Paradigmen, um die Sprachmodelle mit der entsprechenden Grammatik jener Zeit abzugleichen. Mehr als 500.000 Einträge wurden zusätzlich manuell mit bestehenden Wörterbüchern von FineReader verglichen. Grammatikalische Paradigmen und Wortentwicklungen wurden berücksichtigt, um 159 historische Grammatik-Paradigmen, die in den heutigen Sprachmodellen fehlen, hinzuzufügen. Diese Sprachmodelle wurden dann zusammengestellt und an Dokumenten in Frakturschrift getestet.

Um Frakturschrift zu erkennen, haben die Entwicklungsteams von ABBYY spezielle Klassifizierer oder Alphabete erstellt, die Frakturzeichen erkennen können. Das bedeutet, dass für jedes Zeichen durchschnittlich 2.500 Variationen hinterlegt, ein neues Muster-Alphabet angelegt und 31.000 Seiten aus verschiedenen Quellen gesammelt und getestet wurden. Mit einer Vielzahl von Beispieldokumenten bekam die Erkennungs-Engine die Feinabstimmung, um die Besonderheiten des Fraktur-Alphabets wie Ligaturen oder zusammenhängende Buchstaben zu erlernen. Das neue Alphabet wurde dann dem herkömmlichen FineReader System mit einer entsprechenden Oberfläche hinzugefügt und nochmals ausgiebig getestet.

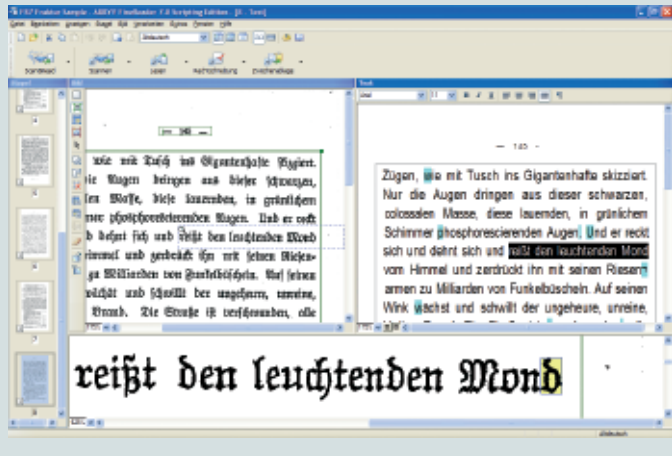
Zusammenarbeit mit großen Archivierungsinstitutionen

Bei der Entwicklung von ABBYY FineReader XIX wurden auch die Bedürfnisse von Universitäten und Forschungseinrichtungen berücksichtigt. Das Produkt wurde in Zusammenarbeit mit dem weltweiten METAE Projekt entwickelt. METAE ist ein Zusammenschluss von Bibliotheken und Digitalisierungsfirmen aus ganz Europa, die zusammen die so genannte METAE Engine entwickeln, um die Umwandlung und Archivierung von historischen Dokumenten wie Büchern, Journalen, Magazinen und

Zeitungen mit Hilfe von Computern zu bewerkstelligen. ABBYY FineReader XIX ist dafür eine Hauptkomponente, um die unschätzbaren historischen Dokumente zu archivieren. Partner in dem METAE Projekt sind unter anderem die Universität von Innsbruck, die Universität von Florenz, die Nationalbibliothek von Frankreich, die Nationalbibliothek von Norwegen, die Friedrich-Ebert-Stiftung, CCS Compact Computer Systeme (Deutschland) und die Cornell Library University (USA).

Wichtigste Features und Funktionen:

ABBYY FineReader XIX basiert auf der OCR Software ABBYY FineReader Corporate Edition. Diese verfügt über eine benutzerfreundliche Oberfläche mit einem Scan&Read Assistenten, der Anwender durch den Prozess des Scannens und der Umwandlung eines Dokuments führt. Einmal umgewandelt, kann der Text einfach bearbeitet und in vielen gängigen Dateiformaten wie Microsoft Word, TXT und durchsuchbaren PDF-Dateien gespeichert werden. Auch das Layout wird übernommen, so dass Spalten, die Position von Bildern und Tabellen wie im Original erscheinen.



Neben den Basisfunktionen von FineReader ist FineReader XIX auch in der Lage, alte Texte mit kunstvollen Schriftarten zu erkennen. Dazu gehören Texte mit schmückenden Ornamenten und romanischen Buchstaben wie das verlängerte „s“, die in frühen englischen und französischen Texten verwandt wurden. FineReader XIX unterstützt folgende Arten von Frakturschrift:

Sprachen:

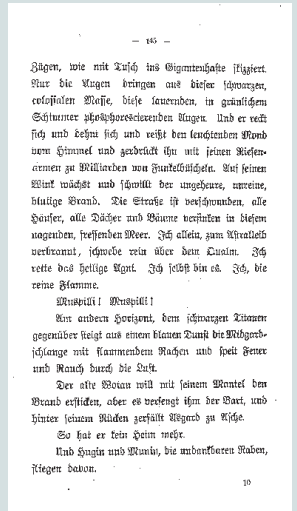
Deutsch, Englisch, Französisch, Italienisch und Spanisch

Schriftarten:

Fraktur, Schwabacher und viele Texturschriften (Gotische Schriften)

Ausgabe-/Speicherformate:

PDF, MS Word (.DOC), MS Excel (.XLS), WordPro, WordPerfect, RTF, HTML, DBF, CSV, TXT und MS Word XML



Zusätzliche Unterstützung für Entwickler:

ABBYY FineReader XIX kann auch mit den Software Entwicklungs-Tools ABBYY FineReader 7.0 Scripting Edition und ABBYY FineReader Engine 7.1 verwendet werden. Entwickler oder Dienstleister können entweder die FineReader Applikation für ihre Bedürfnisse anpassen oder ABBYY OCR für

Frakturschrift in andere Lösungen für Dokumentenarchivierung oder Retrieval. Bei weiteren Fragen bezüglich des Einsatzes von ABBYY FineReader XIX mit den FineReader SDKs kontaktieren Sie bitte das Vertriebsteam von ABBYY Europe unter sales@abbyy.com.

Trial Version:

ABBYY bietet eine vollfunktionsfähige Testversion von FineReader XIX an. Diese Testversion ist zeitlich beschränkt und kann nur eine begrenzte Anzahl von Seiten verarbeiten. Bitte kontaktieren Sie das ABBYY Salesteam, um eine Demoversion zu erhalten.

Spezifikationen

Oberflächensprachen

FineReader XIX ist ein internationales Programm und unterstützt 17 Oberflächensprachen: Deutsch, Englisch, Französisch, Italienisch, Spanisch, Portugiesisch, Niederländisch, Schwedisch, Russisch, Polnisch, Tschechisch, Slowakisch, Ungarisch, Bulgarisch, Ukrainisch, Estnisch und Litauisch.

Systemvoraussetzungen

- PC mit Intel® Pentium®/Celeron®/Xeon™, AMD K6/Athlon™/Duron™ oder kompatibelem Prozessor mit mindestens 200 MHz
- Microsoft Windows 2003, Windows XP, Windows 2000, Windows NT 4.0 (SP6 oder höher), Windows Me/98 (um mit der lokalisierten Oberfläche zu arbeiten, muss die entsprechende Sprache unterstützt werden)
- 64 MB RAM für Windows 2003/XP/2000/NT 4.0; 32 MB RAM für Windows Me/98. Zusätzliche 16 MB RAM werden für jeden Prozessor in einem Multiprozessor-System benötigt.
- 230 MB freier Festplattenspeicher für typische Installation, 70 MB für Programmbetrieb
- 100% TWAIN-kompatibler Scanner, Digitalkamera oder Faxmodem
- Videokarte und Monitor (min. Auflösung 800x600)
- Tastatur, Maus oder anderes Eingabegerät

Unterstützte Formate

Unterstützte Bildformate:

- BMP: schwarzweiß, grau, farbig
- PCX, DCX: schwarzweiß, grau, farbig
- JPEG: grau, farbig
- JPEG 2000, part 1: grau, farbig
- PNG: schwarzweiß, grau, farbig
- TIFF: schwarzweiß, grau, farbig, mehrseitig. Kompressionsmethoden: Unkomprimiert, CCITT Gruppe 3, CCITT Gruppe 3 FAX(2D), CCITT Gruppe 4, PackBits, JPEG, ZIP
- PDF

Unterstützte Speicherformate:

- Microsoft Word Document (*.DOC)
- Rich Text Format (*.RTF)

- Microsoft Word XML Dokument (*.XML) (nur Microsoft Office Word 2003)
- Adobe Acrobat Format (*.PDF)
- HTML. FineReader unterstützt verschiedene Code-Seiten (Windows, DOS, Mac, ISO) und Unicode (UTF-8) Codierung.
- Microsoft PowerPoint Format (*.PPT)
- Comma Separated Values Datei (*.CSV). FineReader unterstützt verschiedene Code-Seiten (Windows, DOS, Mac, ISO) und Unicode (UTF-16, UTF-8) Codierung.
- Nur Text (*.TXT). FineReader unterstützt verschiedene Code-Seiten (Windows, DOS, Mac, ISO) und Unicode (UTF-16, UTF-8) Codierung.
- Microsoft Excel Spreadsheet (*.XLS)
- DBF. DBF. FineReader unterstützt verschiedene Code-Seiten (Windows, DOS, Mac, ISO).

Erkennungssprachen

- 34 Hauptsprachen, für die FineReader Wörterbuchunterstützung und Rechtschreibprüfung anbietet: Deutsch (alte und neue Rechtschreibung), Englisch, Französisch, Italienisch, Spanisch, Portugiesisch, Portugiesisch (Brasilien), Niederländisch, Niederländisch (Belgisch), Schwedisch, Norwegisch (Bokmal), Norwegisch (Nynorsk), Dänisch, Finnisch, Russisch, Estnisch, Lettisch, Litauisch, Polnisch, Tschechisch, Slowakisch, Ungarisch, Bulgarisch, Rumanisch, Kroatisch, Griechisch, Türkisch, Ukrainisch, Armenisch (West, Ost, Grabar), Tatarisch, Katalanisch
- 5 FineReader XIX Sprachen, zur Erkennung von alten europäischen Dokumenten zwischen dem 17.-20. Jahrhundert: Englisch, Französisch, Deutsch, Italienisch und Spanisch
- 133 zusätzliche Sprachen mit lateinischen, kyrillischen oder griechischen Buchstaben (komplette Liste der unterstützten Sprachen unter www.ABBYY.com)
- 4 künstliche Sprachen: Esperanto, Interlingua, Ido und Okzidentalisch
- 6 Programmiersprachen: Basic, C/C++, COBOL, Fortran, JAVA und Pascal.
- Einfache chemische Formeln
- Ziffern

Barcode Types

- 1D: Check Code 39, Check Interleaved 25, Code 128, Code 39, EAN 13, EAN 8, Interleaved 25, CODABAR (ohne Prüfsumme), UCC Code 128, Code 2 of 5 (Industrial, IATA, Matrix), Code 93, UPC-A, UPC-E und Postnet.
- 2D: PDF 417



Weltweit:
ABBYY Software House
P.O. Box 54, Moskau, 129301
Russland
Tel.: +7-095-783-3700
Fax: +7-095-783-2663
E-mail: sales@abbyy.com
Internet: www.ABBYY.com

West-Europa:
ABBYY Europe GmbH
Anglerstr. 6, 80339 München
Deutschland
Tel.: +49-(0)89-511159-0
Fax: +49-(0)89-511159-59
E-mail: sales@abbyy.eu
Internet: www.ABBYY.com

Ost-Europa:
ABBYY Ukraine
P.O. Box 23, 02002 Kiev,
Ukraine
Tel.: +380-44-4909999
Fax: +380-44-4909461
E-mail: sales@abbyy.ua
Internet: www.ABBYY.ua

Nord/Zentral-Amerika:
ABBYY USA
47221 Fremont Boulevard,
Fremont, CA 94538, USA
Tel.: +1-510-2266717
Fax: +1-510-2266069
E-mail: sales@abbyyusa.com
Internet: www.ABBYYUSA.com